Dear Candidate,

Congratulations to have been shortlisted to undertake this challenge statement. This means that you have passed important milestones towards becoming a full-fledged healthcare data scientist.

This challenge is both to evaluate your character, initiative, resourcefulness and commitment towards a data science role in the healthcare sector as well as your current competencies in data science.

The challenge will require some thought and planning to develop an analysis plan and a preliminary analysis with recommendations on future work. Some open-source information is provided to guide you along the path. However, do feel free to extend beyond this information to bring in other information and assumptions that you think will be useful.

The timeline to complete the challenge will be mutually agreed with the hiring manager, but it should not exceed 4 weeks. Scope the project according to the agreed timeline. Do note that there is an optional component in the challenge statement.

Please do not hesitate to contact us if you have any queries and the hiring team will help to clarify the project requirements.

All the best in undertaking this journey with us!

HR Evaluation Committee

Problem Statement

Case Study on Predictive Modelling for Medical Images

Early-stage diagnosis of head and neck cancers, which involve the oral-nasal cavities, pharynx and larynx, is of primary importance in reducing global health burden and patient morbidity (Moccia et al., 2018).[1] Although the standard procedure in diagnosing head and neck cancers in clinical practice is to perform an endoscopy (laryngoscopy) to examine the glottis area and larynx for abnormalities, it is highly dependent on the availability and experience of well-trained endoscopic specialists (Araújo et al., 2019).[2] This is especially apparent in low resource settings (low to middle income countries), where there is a shortage of experienced endoscopists and a lack of accessibility to clinical resources, resulting in missed opportunities of early-stage head and neck cancer screening in patients.

The standard procedure in diagnosing HNC in clinical practice is to use a flexible nasopharyngoscope (FNS) to examine the larynx and other parts of the pharynx for abnormalities. These large systems entail a monitor, image processor, light source, and printer (Figure 1). The FNS procedure takes less than one minute and is easy to learn. The exams are reviewed real time by head and neck surgeons who then help determine clinical assessment and plans.

---

[1] Moccia, S., Vanone, G. O., Momi, E. De, Laborai, A., Guastini, L., Peretti, G., & Mattos, L. S. (2018). Learning-based classification of informative laryngoscopic frames. Computer Methods and Programs in Biomedicine, 158(May), 21–30. https://doi.org/10.1016/j.cmpb.2018.01.030

[2] Araújo, T., Santos, C. P., De Momi, E., & Moccia, S. (2019). Learned and handcrafted features for early-stage laryngeal SCC diagnosis. Medical and Biological Engineering and Computing, 57(12), 2683–2692. https://doi.org/10.1007/s11517-019-02051-5

Figure 1: FNS system at Duke University Medical Center

The Benchmark of Automatic Glottis Segmentation (BAGLS) dataset[3] contains a total of 59,250 endoscopic images with their respective segmentation map together with 559 video snippets representative of images collected with a nasopharyngoscope. Acquired in seven international hospitals, this open source dataset allows the establishment of the training of deep neural networks for disease classification as the metadata contains healthy vs abnormal images derived from the videos. BAGLS dataset can be downloaded from: https://zenodo.org/record/3762320#.Y2nYk3ZBxPY

Using the BAGLS dataset, this problem statement is to classify the images into healthy vs abnormal. The metadata for each image frame contains information such as the age range, gender, disorder status and the parameters of the endoscopic examination. The primary outcome measures for this classification are the:

1. Primary Outcomes:
    a. Area Under the Receiver Operating Curve (AUROC),
    b. Area under the Precision Recall Curve (AUPRC)
    c. F-beta scores, sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV), accuracy
    d. Other accuracy measures where you deem useful
2. Secondary Outcomes:
    a. FLOPs and inferential times across the various DNN architectures

The additional focus for the secondary outcome is on identifying models with similar quality of predictions but significantly less FLOPs and inferential times.

The candidate should be able to propose the following:

(1) **Machine-learning based framework** to deal with the problem (including descriptive, predictive modelling) for the data and predictions for only 2 classes: healthy vs others(or abnormal)

[3] Gómez, Pablo, Kist, Andreas M, Schlegel, Patrick, Berry, David A, Chhetri, Dinesh K, Dürr, Stephan, Echternach, Matthias, Johnson, Aaron M, Kunduk, Melda, Maryin, Youri, Schützenberger, Anne, Verguts, Monique, & Döllinger, Michael. (2020). Benchmark for Automatic Glottis Segmentation (BAGLS) [Data set]. In Scientific Data (1.1a, Vol. 7, Numbers 2052-4463, p. 186). Zenodo. https://doi.org/10.5281/zenodo.3762320

(2) Demonstrate a **preliminary understanding of the complexity of the problem** and being able to translate that into **quick and dirty solution** within the timeframe given
(3) Provide the **necessary software (simulation/wrangling/etc) codes and visualization** to communicate and discuss the insights with stakeholders

Key deliverables:

(1) Presentation deck that is less than 20 slides or less that summarizes the following:
    a. Background of problem and datasets
    b. Methodology, Results, Discussion and Limitations
    c. Conclusion
(2) Codes (in Python/R, or anything that the candidate is familiar with), visualizations for the key variables of concern and the outcomes
(3) Propose promising architectures based on the analysis and list down the limitations of the proposed approach. Proposal on making the predictions explainable for users will be a desired bonus feature [OPTIONAL].

Datasets and Information (find additional open-source information if needed)

- BAGLS dataset can be downloaded from: https://zenodo.org/record/3762320#.Y2nYk3ZBxPY
- The writeup for the BAGLS dataset can be found here: https://www.nature.com/articles/s41597-020-0526-3
- Promising models that have been tested previously are CNN, ResNet, ResNet50, MobileNetV2 and GhostNet
- AUROC and AUPRC of above 80% are achievable for binary classification. Aim to improve on all the outcomes.
- [OPTIONAL] develop an explanatory module if your progress is fast.

Challenge Notes

Please note that as the challenge **does not need to be completely solved** within the agreed time period. The candidate should demonstrate resilience and capability to deal with complex problems, being able to translate and simplify the problem into manageable scope for a quick pilot analysis, plan ahead and understand what it is required to fully address the AI problem. The candidate is encouraged to clarify the problem scope after receiving this problem statement as understanding of the domain is important for clinical problems.

Open-source real-world data is provided for the candidate to demonstrate sufficient competencies in dealing with large scale real-world datasets, understand what is necessary for a problem (that you have scoped), wrangle the correct dataset and utilize that subset of data to propose and develop insights from a preliminary analysis of the data. **Contact the hiring manager for clarifications on the problem statement where necessary.**

Use of real-world datasets and modelling real-world problems are not trivial. Hence, do not attempt to come up with a perfect solution. What we hope to see in the candidate is the ability to work out a **reasonable scope** and carry out what has been proposed in analysis that are sufficiently logical and reasonable. Documentation is important. Hence, put additional explanations in your slides as Annex. Codes should be well documented and explained.

The candidate is reminded to not try to perfect the models or analysis given the limited timeline. We hope the candidate can demonstrate competencies in the various desired areas with reduced scope of the activities appropriately planned and executed.